American Journal of Medical Genetics 63:257–267 (1996)

Computer Assisted Diagnosis of Malformation Syndromes: An Evaluation of Three Databases (LDDB, POSSUM, and SYNDROC)

Jörg Pelz, Volker Arendt, and Jürgen Kunze
Institut für Humangenetik (J.P., V.A., J.K.) and Kinderklinik des Virchow-Klinikum, Berlin (J.K.), Germany

Computer programs which can be used as an aid to diagnose multiple congenital anomaly syndromes have been used for many years, but up to now they have been evaluated very rarely. The diagnostic abilities of three of these systems [LDDB (London Dysmorphology Database), POSSUM (Pictures of Standard Syndromes and Undiagnosed Malformations), and SYNDROC] were analyzed. All three programs are based on an algorithm which defines a diagnosis by a set of phenotypic components all having the same weight (descriptive algorithm). A second algorithm is applied by SYNDROC to rank competing diagnoses in order of probability. This pseudo-Bayesian algorithm provides a coefficient of certitude (CC). For a test the clinical findings of 102 patients who had received a firm diagnosis were used. Two search strategies were tried: "novice's strategy" with all findings taken for a search and "expert's strategy" with a selected set of anomalies. Only those diagnoses that were suggested with the 1st rank, defined as the highest degree of agreement, or the highest CC were studied. The greatest resemblance between suggestions of the databases and the clinical diagnosis was obtained with the expert strategy. The highest number of matches were produced by SYNDROC (80 with expert strategy) and the lowest by POSSUM (54 with novice strategy). The overall agreement between the databases is about 40% for the 1st rank. This number reflects that different authors use different pivotal signs for the description of a syndrome. With the pseudo-Bayesian algorithm 59 cases obtained the highest CC value. Great difficulties exist with the subjective estimates for the calculation of these values; the absolute CC values seem to be meaningless. A small number of unusual cases with special combinations of anomalies provide serious problems for correct diagnosis. © 1996 Wiley-Liss, Inc.

KEY WORDS: CAD, algorithm, databases

INTRODUCTION

Many children are born with malformations. Those who have more than one anomaly pose a special challenge to the attending physician because he must consider whether these signs occur in this patient by coincidence or whether they are part of a syndrome. The clinician faces as a main problem a correct diagnosis, which is a prerequisite for adequate counseling or therapy. The diagnosis of children with malformation syndromes is difficult because there are so many of them and they occur only very rarely, i.e., an individual physician has only limited experience. The rate of new descriptions is increasing steadily and it is not easy to keep pace with this process. In addition, every single syndrome is conceptually polytypic [Beckner, 1959], which means that, at least at the macroscopic clinical level, there is no definition of a set of signs which are both necessary and sufficient to establish a diagnosis.

The word diagnosis has two meanings in medicine: 1) the process of identifying the true (causal) nature of a patient's disorder; and 2) the end product of this process, which serves to put the patient in a classification, because the physician's knowledge of causes, treatment, and prognoses is organized that way. Diagnosis in patients with morphological abnormalities can be obtained in two ways: diagnosis at a glance, where the physician knows the diagnosis, because he has seen a case with the same or very similar characteristics before or the patient presents with a characteristic combination of anomalies which form a typical and distinct

Received for publication January 5, 1996; revision received January 9, 1996.

Address reprint requests to Dr. J. Kunze, Institut für Humangenetik und Kinderklinik des Virchow-Klinikum, Medizinische Fakultät der Humboldt-Universität zu Berlin, Augustenburger Platz 1, 13353 Berlin, Germany.

Dedicated to Jürgen W. Spranger on the occasion of his 65th birthday with admiration and best wishes.

phenotype or diagnosis by an integrative approach, where the physician gathers the clinical data available, evaluates them, and tries to synthesize the findings in order to reach a diagnosis. This process of diagnosis is laborious and the more imaginative aspects of human thinking involved are less well understood.

The different diagnostic tasks and strategies of a syndromologist were described by DiLiberty [1988] and by Aase [1990]. The latter compares the daily work of a syndromologist with the work of those who are engaged in criminology: both look for the tiniest hints and then construct an argument to reach a conclusion/diagnosis. He draws distinctions between three different diagnostic strategies, those of Scotland Yard, of Dr. Watson, and of Sherlock Holmes. "Scotland Yard" stands for an algorithmic approach, the classic, linear/recursive technique. Beginning with the most likely, each diagnostic hypothesis is examined and, if it cannot be sustained, is discarded and another is put forward to be tested in turn. A final diagnosis is achieved when all of the competing candidates can be eliminated.

"Watson's strategy" involves a large part of pattern recognition based on prior experience. The greatest pitfalls of this method lie in the danger of misdiagnosis of conditions with which the practitioner is not familiar, and thus the Dr. Watson model falls short in the diagnosis of rare disorders, which form the overwhelming majority of malformation syndromes.

"Sherlock *Holmes*" starts with a detailed history and an exhaustive physical examination. He selects a few pivotal criteria and these are compared with those found in disorders already familiar to the diagnostician or recorded in the medical literature. If a pivotal sign occurs in many malformation syndromes a long list of possible differential diagnoses will result, which forces the diagnostician to scan a large number of possible disorders; if on the other hand, the most unusual characteristic is selected, the list will be short, but the definite diagnosis may be missed.

If diagnosis at a glance is not possible, much subsequent work consists of searching repetitively lists of signs and syndromes. With the development of fast computer systems the possibility arose of leaving this task to a machine. Research on computer-aided diagnosis in general started in the 1960s [Warner et al., 1964] and in the field of malformation syndromes in the 1980s [Buyse, 1980]. The first programs were developed on main frame computers, with restricted access to a small group of initiated users via modem link; later versions based on personal computers brought these facilities to a broader audience; for a recent review, see Evans [1995].

The present study set out to analyze the diagnostic capabilities of three of these software products. For this test the "London Dysmorphology Database" (LDDB) [Winter et al., 1984, 1991], the "Pictures of Standard Syndromes and Undiagnosed Malformations" (POSSUM) [Marquet, 1987], and "SYNDROC" [Schorderet and Aebischer, 1985] were chosen. Computer programs have been welcomed by syndromologists as valuable sources of references [Stevenson and Hall, 1993], but, at the moment, their usefulness for the clinical practice is unknown because systematic evaluations have been rare. Müller and Grimm [1989] reported one observation with an extended version of LDDB in which they succeeded in narrowing a list of differential diagnoses. Flannery and Peterson [1987] searched 43 clinically undiagnosed patients and made three diagnoses with LDDB. Fryer [1991] reports in a review of POSSUM that he is ". . . aware of a recent case where the diagnoses suggested by POSSUM included a chromosome anomaly which was subsequently detected on reviewing the karyotype. . . ." Stromme [1991] describes the diagnosis of four cases of malformed children with the aid of POSSUM to demonstrate that such a program can be useful to the clinician. He summarizes that in a series of 49 children with multiple malformations, 23 received a diagnosis and that "POSSUM was considered to be useful for diagnostic purposes in 15 of these cases." SYNDROC was evaluated by its authors [Schorderet and Aebischer, 1985; Schorderet, 1987] and yielded a 95% concordance with 100 diagnoses of a clinical geneticist, according to the definition of success by the investigators. Since nobody knows a program better than the authors one can expect that they will gain the best result possible and it is nevertheless essential to test this system in a different setting.

The remark of Wyatt [1991] in discussing computer based knowledge systems ("The LDDB is an example of a medical knowledge base distributed on floppy disk. This details the mode of inheritance and clinical signs of 2,000 congenital malformation syndromes. An editorial panel regularly reviews 1,000 journals and issues annual updates. If malformations interest you, it is difficult to see how the system can be bettered.") is empathetic and enthusiastic but at the moment it is a statement without proof. No one can know whether it is possible to better the system, because it is not known how good it is now. Some specialists underwent the difficult task of compiling the information present in the literature into files for retrieval and surely they did their best, but nobody knows how good their best was. These tools have to be tested and not to be trusted blindly as "state of the art." The results that can be gained by using these systems have to be evaluated as carefully and as objectively as possible.

MATERIAL AND METHODS

Programs

Three programs were used for this study: London Dysmorphology Database Version 1990, Pictures of Standard Syndromes and Undiagnosed Malformations Version 3.0, and SYNDROC Version 4.3. All three systems are organized as relational databases, a model which expresses the data of the real world in one or more tables. Each row of a table represents a syndrome, while each column stands for a trait. If an anomaly is thought to be characteristic for a syndrome a mark is made in that column, which may be simply qualitative (present/absent) or quantitative (probability of this trait given that the disorder is present), depending on the opinion of the authors.

All three programs use a statistical approach to problem solving. A diagnosis is made on a statistical relationship between clinical findings and the specification

of a disorder and not on pathophysiological explanations or on causality. The basic operation applied by all three programs is that of database comparisons and a match is based on the nearest fit rather than on a perfect fit. The anomalies of a patient are compared with the description of all entities in the database. The diagnoses are ranked according to the number of matches gained for every disorder; those with the highest number of fits receive the 1st rank, those with the second highest number of matches the 2nd rank, etc. Diagnoses sharing the same number of anomalies get the same rank, although the combinations of findings may be different for every particular suggested disorder. This procedure was called a descriptive algorithm by Schorderet [1987]. For SYNDROC a second statistical approach has been developed which is based on a concept of Bayesian statistics and has been called a pseudo-Bayesian algorithm [Schorderet and Aebischer, 1985]. Bayes' formula states that P(disease | findings) = [P(findings | disease) × P(disease)]/P(findings). The symbol | has to be read as "given" and P denotes probability. To calculate the most probable diagnosis according to this formula one has to determine three values: P(findings | disease) which is the probability of exactly this set of signs in a specific diagnosis; it is not sufficient to know the frequency of a particular symptom in the disease in question; P(disease) which is the frequency of the disease in the population; its prevalence; and P(findings) which is the probability of the findings in the population.

These values are not generally known and especially P(findings) | disease) is almost impossible to deduce. Schorderet [1992] replaced it with a "subjective estimation of the importance of the sign in relation to the diagnosis" and handled the other two variables in a way not precisely explained. That is why the algorithm is called pseudo-Bayesian. The resulting values are no longer probabilities and are named coefficients of certitude (CC).

Patients

For this study, the medical charts of patients who had visited the Genetic Counseling Unit of the Freie Universität Berlin during the years 1980–1989 for diagnostic purposes were used retrospectively. Eligible were all patients who had received a firm diagnosis for their combination of malformations. They had been worked up by one expert (JK) and his diagnoses were taken as "the gold standard." Excluded were all patients in whom a chromosomal abnormality had been diagnosed, since cases of this cause are not included in the LDDB. From the remaining group the cases for the study were chosen randomly from the data file. If during this selection process the same diagnosis was drawn repeatedly, only the first case was kept and all those who followed with the same diagnosis were disregarded. This procedure was repeated, until a group of 102 patients was selected all of whom had received a different diagnosis.

Analyses

In every case, all the manifestations listed in the medical charts were collected and used for the search in the databases in two different ways:

1) Novice strategy: All signs were used for the diagnostic process (novice traits). This is the method used by a novice in syndromology to approach a diagnosis, because due to limited experience and lacking overview it is difficult for him thus to select symptoms out of the pattern of signs as significant for a diagnosis. The same may be true for the experienced expert, if he has absolutely no idea about the particular case.

2) Expert strategy: Only those signs were used for the diagnostic process, which were judged by an expert (J.K.) to be of diagnostic importance (expert traits).

All patients and their numbers of signs for both search strategies are specified in Table I. In POSSUM and in LDDB the findings are represented in three levels, with the first level as a description of rough clinical regions of the body, the second level as a subclassification of every particular region, and the third level as an expression of the specific anomaly. The authors of LDDB emphasize a strategy of building a search around one or two specific anomalies together with a more general clinical trait [Winter et al., 1984; Winter and Baraitser, 1990], the authors of POSSUM suggest to vary terms in the trait list, introducing new findings, or generalizing the signs with group terms [Marquet, 1991]. Their advice was ignored for this study and all searches were performed exclusively with specific traits. Moreover, care was taken to avoid any bias by seeking a diagnosis with inappropriate descriptions of the anomalies in the following way: Since each diagnosis was known in advance, the lists of manifestations of every syndrome were looked up in all databases. The descriptions of our patients were modified as far as possible to reach the highest number of fits with the specific traits for the respective database; those anomalies, that were not part of the description of a syndrome in a database were utilized, unchanged, for each analysis. With this list, that was optimally adapted for producing the diagnosis in question with the respective database, each trial was carried out. All results have to be assessed, taking this procedure into account. It guarantees the optimal attainable rank for each case with the given set of abnormalities. Success was defined as the suggestion of the clinical diagnosis with the 1st rank for the descriptive algorithm and with the highest CC for the pseudo-Bayesian algorithm.

RESULTS

The patients of this study were described using an average of 7.1 signs (range 1 to 15, interquartile range 5 to 9). The average number of expert signs was 4.6 (range 1 to 12), more than 80% of all patients had between three and seven anomalies.

Only three clinical cases were not described in any of the three databases and there was no case that was not described at least in one database. Of the 102 syndromes, both LDDB and SYNDROC did not include a description of one, whereas POSSUM did not contain a representation of three. For those entities that were present, the number of traits for the descriptions of a specific disorder showed large differences between the

TABLE I. List of All Cases

Number	Clinical diagnosis (syndrome, if not stated otherwise)	McKusick	Sex	Age	Number of novice's traits	Number of expert's traits
1	Biedl-Bardet	209900	f	14 y	5	4
2	Goldenhar symptome complex	164210	m	1 m	12	7
3	Smith-Lemli-Opitz	270400	m	8 m	14	10
4	Wiedemann-Beckwith	130650	f	1 m	5	5
5	Pierre-Robin-Sequence	311900	f	1 m	4	2
6	Klippel-Feil anomaly	148900	m	11 m	7	3
7	Fetal alcohol	—	f	1 y	10	4
8	Franceschetti	154500	m	17 d	5	5
9	Acrodysostosis	101800	f	21 y	15	6
10	Cerebro-oculo-facio-skeletal	214150	m	2 m	9	5
11	Incontinentia pigmenti	308300	f	31 y	5	4
12	Short-rib-polydactyly (Majewski)	263520	f	4 d	6	2
13	Spondylo-epiphyseal dysplasia congenita	120140	m	3 y	9	5
14	Otopalatodigital (Type 2)	304120	m	1 d	10	6
15	Contractural arachnodactyly (Beals)	121050	m	17 y	12	6
16	IVIC	147750	m	4 m	3	3
17	Hypomelanosis Ito	146150	m	2 y	4	3
18	Ectodermal dysplasia	305100	f	4 y	15	7
19	Holt-Oram	142900	f	5 y	4	3
20	Pfeiffer	101600	m	17 y	8	4
21	Stilling-Türk-Duane	126800	m	3 y	5	2
22	Waardenburg	193500	m	8 m	9	7
23	Ectrodactyly-ectodermal-dysplasia-clefting	129900	m	31 y	6	3
24	Thrombocytopenia-absent-radius	274000	m	7 y	13	4
25	Distal arthrogryposis	108120	f	24 y	2	2
26	Oro-acral	157900	m	2 m	11	6
27	Ehlers-Danlos	130000	m	23 y	8	5
28	Campomelic dysplasia	211970	f	1 d	10	5
29	Spondylo-thoracal dysplasia (Jarcho-Levin)	277300	m	2 m	6	4
30	Cornelia de Lange	122470	m	2 y	7	7
31	Marfan	154700	f	12 y	8	8
32	Hypochondroplasia	146000	f	28 y	6	5
33	Achondroplasia	100800	f	3 y	14	12
34	Larsen	150250	f	1 y	10	9
35	Angelman	234400	f	10 y	6	5
36	Pena-Shokeir	214150	f	1 y	9	6
37	Prader-Willi	176270	m	7 m	8	4
38	Otopalatodigital (Type 1)	311300	m	6 y	9	6
39	Greig's cephalopolysyndactyly	175700	f	1 y	8	6
40	Myotonic dystrophy (Curshmann-Steinert)	169000	m	30 y	6	2
41	Silver-Russell	312780	m	2 m	9	7
42	VACTERL association	192350	?	Stillbirth	10	8
43	Progeroid (Wiedemann-Rautenstrauch)	176670	f	1 m	11	7
44	Adams-Oliver	100300	f	1 m	10	3
45	Dyschondrosteosis Léri-Weill	127300	m	38 y	4	4
46	Walker-Warburg	236670	f	2 d	7	2
47	Jeune	208500	m	14 m	2	2
48	Albinismus	203100	f	4 m	3	2
49	Seckel	210600	f	3 y	6	4
50	Multiple-pterygium (Escobar)	265000	f	31 y	7	4
51	LADD	149730	f	1 y	14	5
52	Williams-Beuren	194050	f	9 y	10	7
53	Stickler	108300	f	7 y	6	4
54	Prune belly sequence	100100	m	2 y	9	4
55	Noonan	163950	f	1 y	8	3
56	Acrocephalosyndactyly	101200	m	5 d	5	4
57	Coffin-Lowry	306300	f	12 y	10	6
58	CHARGE association	214800	m	2 m	9	5
59	Ellis-van Creveld	225500	f	1 m	4	3
60	Fronto-facio-nasal	229400	f	1 y	4	4
61	Rubinstein-Taybi	180849	m	6 m	9	3
62	Goltz-Gorlin	109400	f	11 d	10	9
63	Thanatophoric dysplasia	187600	f	27 y	11	6
64	Ivemark	208530	f	2 m	4	4
65	Cerebral gigantism	117550	f	8 m	5	2
66	Holoprosencephaly	236100	m	7 d	9	5

TABLE I. List of All Cases (*continued*)

Number	Clinical diagnosis (syndrome, if not stated otherwise)	McKusick	Sex	Age	Number of novice's traits	Number of expert's traits
67	Klippel-Trenaunay	149000	m	6 y	3	3
68	Whistling-face	193700	m	16 m	7	3
69	Diastrophic dysplasia	222600	m	13 m	11	6
70	Miller-Dieker	247200	f	2 y	13	6
71	Oculo-dento-digital	164200	f	1 y	9	6
72	Coffin-Siris	135900	m	?	8	7
73	Cockayne	216400	m	17 y	11	8
74	Acro-renal polytopic defect	102520	m	10 m	5	2
75	De-Barsy	215150	m	2 y	8	4
76	Trichorhinophalangeal	190350	m	32 y	4	4
77	Aicardi	304050	f	2 m	5	5
78	Lowe	309000	m	7 y	4	4
79	Sirenomelia	—	?	1 d	7	7
80	Cenani-Lenz	212780	f	15 y	5	5
81	Marinesco-Sjögren	248800	f	5 y	5	4
82	Chondrodysplasia punctata	215100	f	Stillbirth	5	4
83	Menkes	309400	m	?	2	2
84	BBB	145410	f	30 y	5	2
85	Meckel	249000	m	1 d	11	6
86	Gardner	175100	m	31 y	1	1
87	Dysostosis cleidocranialis	119600	f	10 m	4	4
88	Osteogenesis imperfecta	120150	f	26 y	3	3
89	Proteus	176920	f	6 y	4	3
90	Robinow	180700	f	4 m	3	2
91	Rett	312750	f	3 y	3	3
92	Potter sequence	173900	m	16 m	2	2
93	Roberts	268300	?	Stillbirth	5	3
94	Poland anomaly	173800	m	?	2	2
95	Bartter	241200	m	3 y	7	3
96	Sturge-Weber	185300	f	35 y	2	1
97	Crouzon	123500	f	5 y	9	4
98	Schinzel-Giedion	181450	f	1 d	7	4
99	Fraser	219000	m	1 m	9	6
100	Opitz-Trigonocephaly	211750	m	3 m	7	7
101	Black Lock-Albinism-Deafness[a]	227010	m	6 d	2	3
102	Genée-Wiedemann	263750	f	18 y	12	12

[a] This case was described as a new syndrome [Groß et al., 1995], but it shares all clinical signs that are characteristic for the Black Lock-Albinism-Deafness syndrome.

databases. The average number used in LDDB was 20 (range 2 to 36), in POSSUM 41 (range 3 to 94), and in SYNDROC 18 (range 4 to 48).

The master lists for the description of the morphological abnormalities of the three databases contain different numbers of anomalies. The lowest number of 890 is found in POSSUM, while SYNDROC and LDDB give both about 1100. Not all signs of the respective clinical case were found in the master lists of the databases. Of the 7.1 average clinical abnormalities per patient, 7.0 equivalent descriptions could be used for a search in POSSUM, 6.5 in LDDB and 6.1 in SYNDROC. The smallest list of phenotypic components gave maximal possibilities for the description of the patients for a search with the database; for an overview see Table II.

The main purpose of the databases is decision support in diagnosis. The reliability of the databases was estimated by the criterium that they gave the correct diagnosis with the first rank, be it alone or among others. The results are given in Table III.

With the expert traits the clinical diagnosis is suggested by LDDB in 68%, by POSSUM in 63% and by

TABLE II. Number of All Traits in the Databases and of the Number of Equivalent Signs for the Description of Cases

	Number of abnormalities of the clinical case	LDDB	POSSUM	SYNDROC
Traits in database	—	1087	890	1112
Novice's traits	7.1	6.5	7.0	6.1
Expert's traits	4.6	4.4	4.6	4.0

TABLE III. Number of Correctly Suggested Diagnoses by Utilization of Descriptive Algorithm*

	LDDB	POSSUM	SYNDROC
Novice's strategy			
1. rank	64 (31)	54 (22)	70 (49)
2. rank	19	24	14
3. rank	9	11	9
Higher rank	9	10	8
Not described	1	3	1
Expert's strategy			
1. rank	69 (30)	64 (18)	80 (40)
2. rank	19	24	13
3. rank	9	8	5
Higher rank	4	3	3
Not described	1	3	1
Within same database			
With both strategies	60	52	67
Only with expert's strategy	9	12	13
Only with novice's strategy	6	2	3
With neither strategy	27	36	19

* The numbers in parenthesis represent the number of cases, in which the correct diagnosis was suggested as the only one.

SYNDROC in 78%, the results after deliberate utilization of all clinical symptoms are worse for all three systems.

It should be noted, that for all three databases the highest number of correct diagnoses without concurring differential diagnoses (49) were yielded with SYNDROC using a novice strategy and the lowest number (18) was produced with POSSUM using expert strategy. These numbers are given in parentheses in Table III. The application of both strategies within one database, which may simulate the everyday situation in a clinical setting, produces the following results: For LDDB, 60 diagnoses are suggested with both strategies, nine only using expert traits and six only with novice traits. With POSSUM, 52 disorders are diagnosed from both groups of traits, 14 additional diagnoses come only with one search strategy, 12 entering expert abnormalities, and 2 while searching with those of a novice. SYNDROC gives the correct identical diagnoses of 67 cases by utilization of both strategies and 16 additional with one of both sets of traits, 13 with expert traits, and 3 with novice signs.

Clinical cases that were not suggested with the 1st rank are to be expected with a higher (worse) rank in the extended differential diagnosis, if the description of the database and of the clinical case had at least one anomaly in common. Using the novice strategy, between 17 (SYNDROC) and 21 (POSSUM) cases were suggested with the third or a higher rank, while with the expert strategy there were between 8 (SYNDROC) and 13 (LDDB).

The results for running 2 or 3 databases at a time are shown in Figures 1 and 2 for both search strategies.

Expert traits. The highest resemblance is found for SYNDROC and LDDB (59), the lowest for LDDB and POSSUM (50), while SYNDROC and POSSUM (54) give intermediate results. The application of SYNDROC to any of the two other databases fails to give 12 diagnoses at all, while this number is 19 for the use of LDDB and POSSUM together.

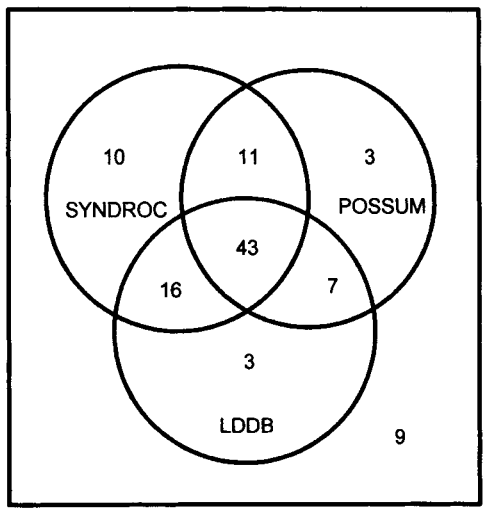

Fig. 1. Number of syndromes, which have been correctly suggested with the 1st rank by the different databases using expert's strategy and the descriptive algorithm.

Novice traits. A very similar picture emerges with this strategy, resulting in 54 diagnoses alike with SYNDROC and LDDB, 47 with LDDB and POSSUM, and 46 with POSSUM and SYNDROC. The utilization of SYNDROC and LDDB does not produce 20 diagnoses, while for POSSUM with LDDB or with SYNDROC the failures are 24 and 29, respectively. Only 43 identical first ranked diagnoses are received with all three databases using expert traits for the computer search and 41 with novice traits. The former strategy does not produce 9 diagnoses at all, while the latter fails to obtain 18.

The nine cases, which were not suggested with a first rank diagnosis with any of the search strategies were analyzed using seven monographs of the standard syndrome literature. For a comparison of the description in

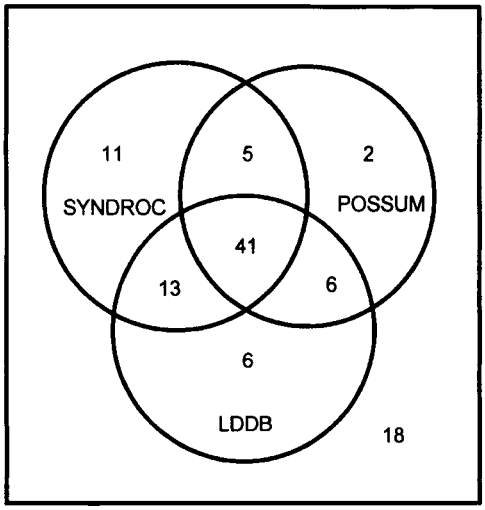

Fig. 2. Number of syndromes, which have been correctly suggested with the 1st rank by the different databases using novice's strategy and the descriptive algorithm.

the databases and by our expert, all anomalies were listed that were considered important for the establishment of the respective diagnosis by the authors or that were mentioned as characteristic in more than 50% of the monographs which contained the description. The results are summarized in Table IV.

Two of the clinical cases (number 95 and 101) were described in only one of the databases, so that the probability of finding them with a search was seriously compromised; descriptions of these syndromes were only found in four and two of the monographs, respectively. For the seven remaining cases, at least four of the expert traits were found in more than half of the literature for the syndromes number 7, 13, and 34. Case number 7 (all three databases) and case number 13 (SYNDROC and LDDB) are not described in accordance with the literature in the databases. For case number 34, of the 9 expert traits 4 to 6 were also used in the databases, but in every program other syndromes existed, whose descriptions matched with a higher number of the anomalies. For the remaining four cases the combinations of findings in the patients differ from those of the databases and the literature as well.

SYNDROC, in using the pseudo-Bayesian algorithm, weighs the importance of all abnormalities with a supposed diagnosis. In doing so a CC-value is estimated for every possible syndrome in the database. These CC-values are shown in Figure 3 for all 102 cases and for both search strategies. The CC-values were truncated after the first decimal digit. With the expert traits 17 cases got a CC-value of less than 0.1, the mean for all cases was 0.28 and more than 90% of all diagnoses received a CC-value of less than 0.6. Of all 102 cases 59 gained the highest CC-values of all diagnoses produced within a search, in 43 cases a higher value was calculated for a different syndrome.

For 52 patients the highest CC-value was yielded both with the expert traits as well as with the novice traits; 6 clinical cases got the highest CC-value exclusively when the novice strategy was applied, 7 when the expert strategy was tried, and in 37 cases another diagnosis got a higher estimate.

An overview of the combined application of the pseudo-Bayesian algorithm and the descriptive algorithm with SYNDROC is given in Figure 4.

Expert strategy. Both algorithms yielded an identical result for 52 cases, 28 diagnoses were suggested with the 1st rank exclusively by the descriptive algorithm, 7 with the highest CC-value only by the pseudo-Bayesian algorithm; 15 diagnoses were not obtained with either algorithm.

TABLE IV. Nine Syndromes Which Were Not Suggested by Any of the Three Databases With the First Rank Using the Expert-Signs and Their Descriptions in the Literature*

Syndrome number	7	11	20	26	34	38	41	95	101
Syndrome described in LDDB	+	+	+	+	+	+	+	–	+
Syndrome described in POSSUM	+	+	+	+	+	+	+	–	–
Syndrome described in SYNDROC	+	+	+	+	+	+	+	+	–
Number of descriptions in literature	7	5	7	7	6	7	7	4	2
Number of traits in literature	25	19	25	25	23	27	25	12	9
Clinical expert-signs	4	5	4	6	9	6	7	3	2
Clinical expert-signs in >50% of the literature	4	4	1	0	4	2	2	1	1
Expert-signs in LDDB	2	3	1	2	4	1	3	–	0
Expert-signs in POSSUM	2	4	0	5	6	2	4	–	–
Expert-signs in SYNDROC	0	1	2	2	4	0	3	1	–
Clinical novice-signs	10	9	8	11	10	9	9	6	2
Clinical novice-signs in >50% of the literature	6	4	7	9	6	6	6	3	1
Novice-signs in LDDB	2	4	2	5	4	1	4	–	0
Novice-signs in POSSUM	4	6	2	7	6	2	5	–	–
Novice-signs in SYNDROC	2	2	2	4	4	1	4	4	–

*Literature used for comparison: Burg et al. [1990], Buyse [1990], Goodman and Gorlin [1983], Gorlin et al. [1990], McKusick [1994], Wiedemann et al. [1989], Witkowski et al. [1991].

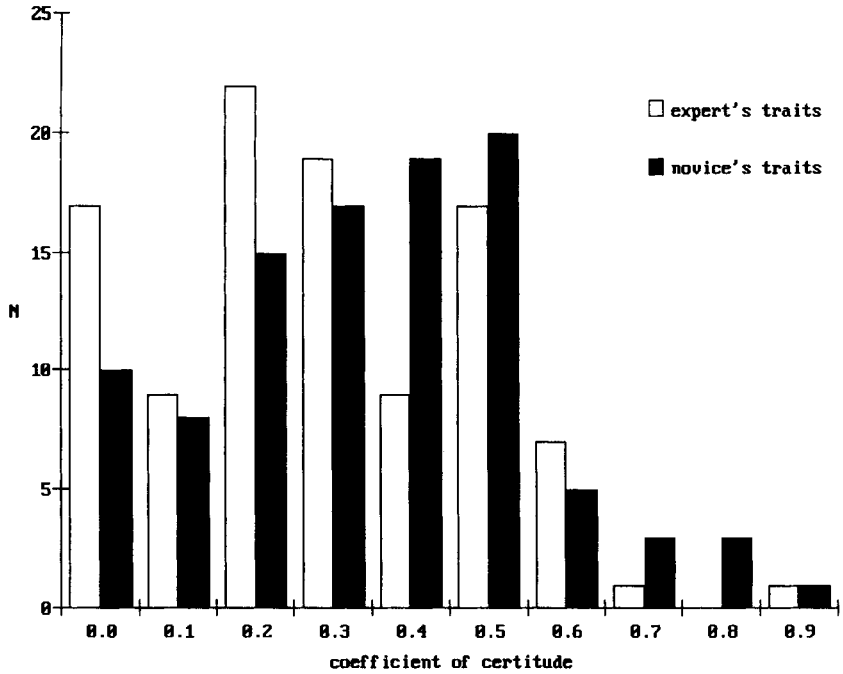

Fig. 3. Frequencies of the coefficients of certitude obtained by using expert's traits and novice's traits with the pseudo-Bayesian algorithm.

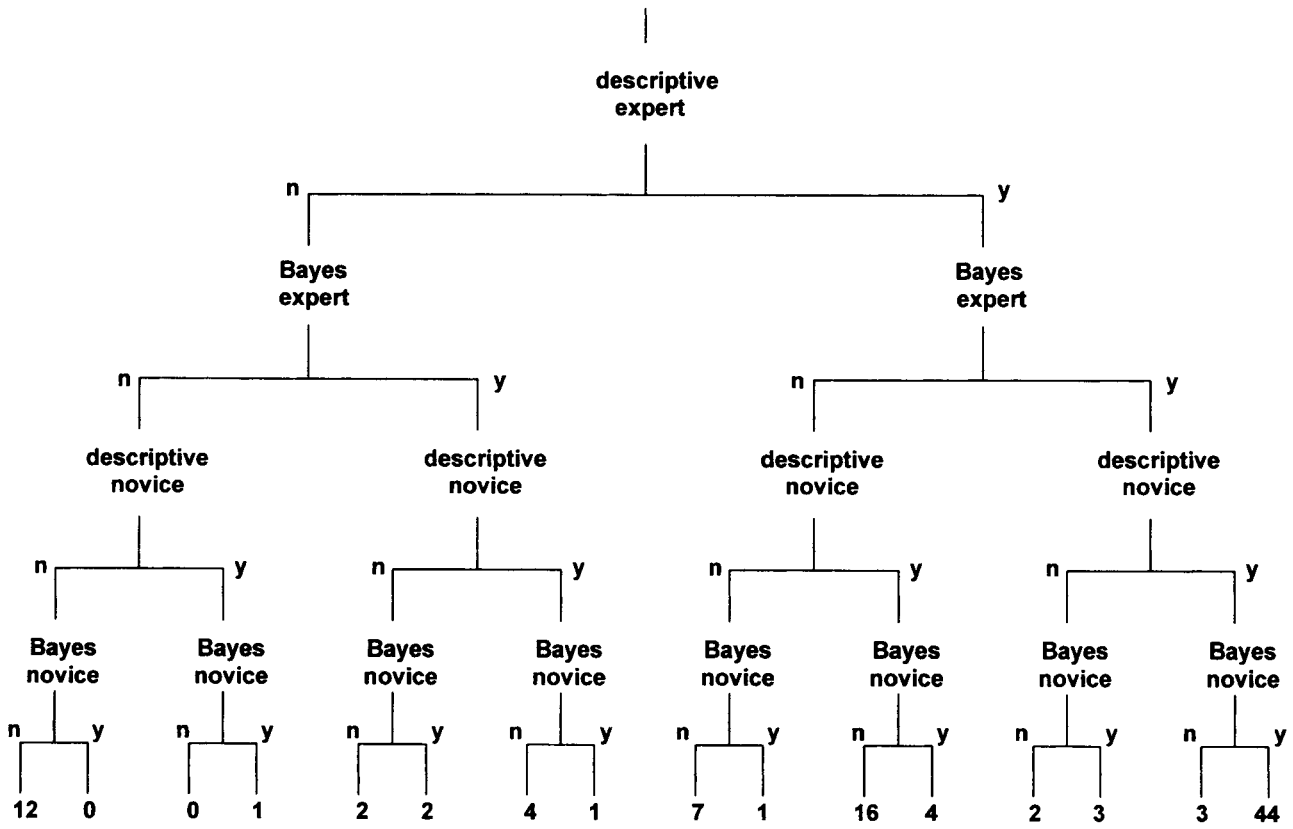

Fig. 4. Number of correctly suggested diagnoses by using SYNDROC with both algorithms and both search strategies (y, correct diagnosis suggested; n, correct diagnosis not suggested).

Novice strategy. A concordant diagnosis was proposed in 49 cases, 21 diagnoses were returned by the descriptive algorithm, and 9 by the pseudo-Bayesian algorithm alone; 23 cases did not match with either algorithm.

DISCUSSION

Computer programs like the ones analyzed here are designed to deal with complex problems. Their purpose is to support decisions of physicians. For the evaluation process of such systems one has to differentiate between the program as a software product and as a model for the process it has been designed for. Only the latter is the subject of this study. For the former there is no need for testing, since all three programs at issue here are constructed in a user-friendly manner. The communication with these programs is menu driven and when the clinical abnormalities have been entered, the time for the performance of a search is very short. Nevertheless, experts seem to be reluctant to use computer tools. Although it is possible to search a database for combinations of malformations very easily, in a review of 100 newly described syndromes it was found, that only in 16 cases had the authors made use of syndrome databases (Stich et al., submitted). One way to overcome these attitudes is a repeated analysis whether a program does what is required [Lundsgaarde, 1987] and whether it is working correctly according to the needs of the practitioners [O'Keefe et al., 1987], so that the potential user knows in advance what he can expect.

The first problem for the user after having finished the history and the data gathering is to enter his findings into the program. For this purpose the programs provide master lists, which cannot be changed by the user. These lists represent the interpretation of the authors of how, and how many, clinical traits have to be described for the representation of malformation syndromes and the user has to cope with this. This study demonstrated that the length of the list did not correspond to the possibility of describing clinical findings. The terminology used by POSSUM seems to be more similar to the one used by our study group than that of the other two programs; a finding which cannot be generalized and which may be quite different for other specialists.

The second problem for the user is how many and which of his findings to utilize for the diagnostic process. This will depend on his experience, which has a great influence on a clinician's judgment of the importance of an abnormal finding on the probability of a diagnosis [Balla, 1982]. Taking for granted that the generation of hypotheses and the deduction and testing of consequences (hypothetico-deductive process) is an effective scientific method, it has been shown that experts produce hypothesis earlier during the work up of a patient and that their hypotheses are more often correct, than those created by a novice [Leaper et al., 1973; Balla, 1982]. To get an impression of the influence of the specificity of manifestations used for a search, the strategies of taking all anomalies and of using a set of traits defined by an expert were used. The results of these strategies depend on the methods of data processing in a given program.

This is one of the main difficulties which authors of a decision support system have to conceptualize. The easiest procedure is to add all fits between the anomalies of the clinical case and the set of traits of all diagnoses of the database. Spranger [1978] criticized this additive method (descriptive algorithm) for making a diagnosis by simply comparing the degree of quantitative agreement. He argued that different anomalies have varying diagnostic importance, primary and secondary malformations receive the same weight, and the impression of the "Gestalt" is lost using this method. He cites as an example that a list of anomalies for trisomy 13 and for the Meckel-syndrome may be similar, but one glance at the face of patients with these disorders will show that they are different.

Spranger's last argument will not pose any insoluble problem, since the differential diagnosis of two equally probable syndromes will be easy to decide. His first two arguments are in favor of a Bayesian approach. The ability to attach correct weights to findings is a major skill of an experienced clinician. The way this is achieved in problem solving in clinicopathological conferences has been investigated but is still not understood [Eddy and Clanton, 1982]. For the tested databases only SYNDROC provided weights for traits. Since all values for the formula of Bayes are unknown and had to be replaced by subjective estimates, this procedure was called pseudo-Bayesian algorithm.

The implementations of both models for deducing a diagnosis have been tested with both search strategies. For the descriptive algorithm, overall the best results were obtained with SYNDROC. By the use of expert traits a higher rate of correctly suggested diagnoses was obtained for all databases. The gain in absolute numbers amounted to 10 for POSSUM and SYNDROC and to 5 for LDDB. The reason for this is that the expert picks out those anomalies of a patient which are specific and characteristic for his disorder and that the same abnormalities have been stored in the databases. The utilization of all findings with the descriptive algorithm dilutes in a certain way the description of the disorder. Some of the additional signs may be less characteristic for the syndrome at hand but may lead to an overvaluation of very different entities, for which the additional anomalies may be specific. For SYNDROC our results are better than those reported by Schorderet [1987], who received the correct diagnosis in 54% in the 1st rank.

The results with the pseudo-Bayesian algorithm for the highest CC, 58% and 57% correct diagnoses with expert and novice traits, respectively, are also slightly better than those of Schorderet's study (55%). If one applies less strict criteria and takes into consideration the second and the third highest CC value, a 72% success rate is yielded by application of expert traits and 75% success with novice traits. These values do not reach the same magnitude as those of Schorderet [1987], who obtained 92%. At the time when Schorderet did his analysis, the number of syndromes in SYNDROC was

much smaller. It was comparable in this respect to a system of Wiener et al. [1987], who got 91% correct diagnoses with a modified Bayesian method as well. The broadening of the database seems to be responsible for the reduced diagnostic accuracy.

Striking in the present study is the magnitude of the CC values. Although Schorderet [1992] advised to look more for the values of competing diagnoses than for the absolute value, it is not easy to understand why only a very small proportion of cases got a CC of higher than .5, and, moreover, that 15 cases tested with expert strategy and 10 with novice strategy obtained 0.0. Taking into account that all cases had a firm diagnosis, it seems to be necessary to change some of the subjective estimates that went into the pseudo-Bayesian formula.

The third problem for the user of a decision support system is to decide what to do with the results. Surely, none of the systems tested makes a diagnosis but they do provide some suggestions, which have then to be evaluated by the diagnostician. How many differential diagnoses should be taken into account? Winter and Baraitser [1990] and Aase [1990] are in favor of a small list, less than 10, while Schorderet [1991] suggested that a high number of differential diagnoses may be attractive for an expert, who could easily pick out any relevant diagnosis. For the present study very strict criteria were applied in taking only 1st ranked diagnoses and those with the highest CC-values. This seems to be justified, because a search strategy was used that guaranteed the best possible result. With less stringent criteria, e.g., taking all diagnoses up to the 3rd rank, all three databases produce a success rate of about 90% with novice strategy and more than 95% with expert strategy. The application of strict criteria shows that the overall agreement between one expert and the three databases is about 40% for the 1st rank. This demonstrates that experts, who agree on a clinical diagnosis, do this with different pivotal signs in mind and further it clearly shows that a dilemma exists if weights of different traits for a Bayesian algorithm have to be estimated.

One weakness of the databases was uncovered by the analysis of those cases that were not suggested by any of them with the descriptive algorithm. There is a small number of syndrome descriptions which is not in agreement with the majority of the literature. These few cases can be corrected and the more such cases will be found with intensive tests of the programs, the better the databases will become. A more serious issue is that of the unusual cases, which clearly belong to a particular syndrome, but show a special combination of anomalies. It is improbable that an extension of the description in the database will provide better results. The phenotypic characteristics become diluted if the additive model is applied and, for a Bayesian model, the experience with these cases is too sparse to estimate any meaningful values for the calculation of probabilities. With this class of challenges, where a high level of support is desired, the expert is left alone and will have to deal himself. Manning [1987] explored in her "why Sherlock Holmes can't be replaced by an expert system" that programs are not able to pose problems well and cannot evaluate the importance of data, when these are unexpected, while the skilled human expert can distinguish between data which stand in need of an explanation and data which would support one explanation over another.

ACKNOWLEDGMENTS

We thank Dr. Martin Digweed for helpful discussions and many suggestions.

REFERENCES

Aase JM (1990): "Diagnostic Dysmorphology." New York: Plenum Medical Book Company.

Balla JI (1982): The use of critical cues and prior probability in decision making. Methods Inf Med 21:9–14.

Beckner M (1959): "The Biological Way of Thought." New York: Columbia University Press.

Burg G, Kunze J, Pongratz D, Scheurlen PG, Schinzel A, Spranger J (1990): "Leiber—Die klinischen Syndrome." München: Urban and Schwarzenberg.

Buyse ML (1980): Center for birth defects information services. BD:OAS XVI (5):83–91.

Buyse ML (1990): "Birth Defects Encyclopedia." Dover: Center for Birth Defects Information Services.

DiLiberti JH (1989): Use of computers in dysmorphology. J Med Genet 25:445–453.

Eddy DM, Clanton CH (19820: The art of diagnosis. Solving the clinipathological exercise. N Engl J Med 306:1263–1268.

Evans CD (1995): Computer systems in dysmorphology. Clin Dysmorphol 4:185–201.

Flannery DB, Peterson SL (1987): Practice analysis of dysmorphology diagnostic data bases. Am J Hum Genet 41:A49.

Fryer A (1991): POSSUM (Pictures of Standard Syndromes and Undiagnosed Malformations). J Med Genet 28:66–67.

Goodman RM, Gorlin RJ (1983): "The Malformed Infant and Child." New York: Oxford University Press.

Gorlin RL, Cohen MM, Levin LS (1990): "Syndromes of the Head and Neck." New York: Oxford University Press.

Groß A, Kunze J, Stoltenburg-Didinger G, Grimmer I, Maier RF, Obladen M (1995): A new syndrome: an autosomal-recessive neural crest syndrome with albinism, black lock, cell migration disorder of the neurons of the gut and deafness: ABCD syndrome. Am J Med Genet 56:322–326.

Leaper DJ, Gill PW, Staniland JR, Horrocks JC, De Dombal FT (1973): Clinical diagnostic process: an analysis. Br Med J *iii*: 569–574.

Lundsgaarde HP (1987): Evaluating medical expert systems. Soc Sci Med 24:805–819.

Manning RC (1987): Why Sherlock Holmes can't be replaced by an expert system. Phil Stud 51:19–28.

Marquet C (1987): "P.O.S.S.U.M. User's Manual." Melbourne: C.P. Export Pty Ltd.

Marquet C (1991): "P.O.S.S.U.M. User's Manual." Fourth Edition. Melbourne: C.P. Export Pty Ltd.

McKusick VA (1994): "Mendelian Inheritance in Man." Baltimore: The Johns Hopkins University Press.

Müller B, Grimm T (1989): Computerunterstützte Syndromdiagnostik. Ein Beispiel. Pädiatr Prax 39:575–579.

O'Keefe R, Balci O, Smith E (1987): Validating expert system performance. IEEE Expert 2:81–89.

Schorderet DF (1987): Diagnosing human malformation patterns with a microcomputer: Evaluation of two different algorithms. Am J Med Genet 28:337–344.

Schorderet DF (1991): Using OMIM (On-line Mendelian Inheritance in Man) as an expert system in medical genetics. Am J Med Genet 39:278–284.

Schorderet DF (1992): "Diagnosing Human Malformation Patterns with SYNDROC." User's manual, version 4.3.

Schorderet D, Aebischer P (1985): SYNDROC: Microcomputer based differential diagnosis of malformation patterns. Arch Dis Child 60:248–251.

Spranger J (1978): Syndrome: Klinische Erfassung und Bedeutung. Mschr Kinderheilk 126:252–258.

Stevenson RE, Hall JG (1993): Acknowledgments. In: Stevenson RE, Hall JG, Goodman RM (eds): "Human Malformations and Related Anomalies." Vol 1. New York: Oxford University Press, p xi.

Stromme P (1991): The diagnosis of syndromes by use of a dysmorphology database. Acta Paediatr Scand 80:106–109.

Warner H, Toronto A, Veasy L (1964): Experience with Bayes' theorem for computer diagnosis of congenital heart disease. Ann NY Acad Sci 115:2–16.

Wiedemann H-R, Kunze J, Dibbern H (1989): "Atlas der Klinischen Syndrome." Stuttgart: Schattauer.

Wiener F, Gabbai M, Jaffe M (1987): Computerized classification of congenital malformations using a modified Bayesian approach. Comput Biol Med 17:259–267.

Winter RM (1991): Computers in dysmorphology: The London Dysmorphology Database. Am J Hum Genet 49:A28.

Winter RM, Baraitser M (1990): "London Dysmorphology Database." Oxford: Oxford University Press.

Winter RM, Baraitser M, Douglas JM (1984): A computerised data base for the diagnosis of rare dysmorphic syndromes. J Med Genet 21:121–123.

Witkowski R, Prokop O, Ullrich E (1991): "Wörterbuch für die genetische Familienberatung." Berlin: Akademie Verlag.

Wyatt J (1991): Computer-based knowledge systems. Lancet 338:1431–1436.